Multimedia Tools and Applications manuscript No.
(will be inserted by the editor)

An experimental evaluation of ontology-based user profiles

Frank Hopfgartner · Joemon M. Jose

Authors' version of manuscript published by Springer (DOI 10.1007/s11042-012-1254-2)

Abstract In recent years, a number of research works have been carried out to improve the information retrieval process by exploiting external knowledge, e.g., by employing ontologies. Even though ontologies seem to be a promising technique to improve the retrieval process, hardly any study has been performed to evaluate the use of ontologies over a longer time period to model user interests. In this work we introduce an ontology based video recommender system that exploits implicit relevance feedback to capture users' evolving information needs. The system exploits a generic ontology to organise users' interests. We evaluate the recommendations by performing a user-centred multiple time-series study where participants were asked to include the system into their daily news gathering routine. The results of this study suggest that the system can be successfully employed to improve personal information seeking tasks in news domain.

Keywords Video Retrieval · Multiple time series study · Personalisation

1 Introduction

An analyis of state-the-art research within TRECVid indicates that content-based video retrieval performance is still far away from their textual counterparts [8]. An interesting approach to narrow this performance gap is to further enrich video documents using external data sources. Fernández et al. [11], for instance, have shown that ontology-based search models can outperform classical information retrieval models at a web scale. The advantage of these models is that external knowledge is used to set the content into their semantic context. In this paper, we study how such external knowledge can be exploited to recommend news videos that match users' personal interests. We thus build this research upon previous work presented in [15], where we evaluated a news video recommender system that exploits ontologies to set news stories into their

Frank Hopfgartner
International Computer Science Institute, Berkeley, CA, USA
E-mail: fh@icsi.berkeley.edu

Joemon M. Jose
University of Glasgow, Glasgow, UK
E-mail: joemon.jose@glasgow.ac.uk

semantic context using a simulation-based evaluation scheme. The objective of this study is to verify these findings from the users' perspective. Hence, we conducted a long-term user-centred study.

In [15], we introduced a novel news video recommender system which captures daily broadcasting news, and segments the bulletins into semantically related news stories. DBpedia is a structured representation of Wikipedia [2] and hence is exploited to set these stories into their context. This semantic augmentation of news stories is used as the backbone of our news video recommendation. We introduced an implicit user modeling approach which automatically captured users' evolving information needs, representing interests in a dynamic user profile. In our simulation based evaluation we aimed to verify two research questions. The first question was whether implicit relevance feedback can be used to create long-term user profiles. The second research question was to study whether a generic ontology can be exploited for accurate news video recommendations. For this, we exploited DBpedia to place the news stories into their semantic context. In order to evaluate these research questions, we proposed a simulation-based evaluation methodology which models user interactions by mimicking users interacting with the video recommender system.

A simulation-based evaluation evaluation scheme is needed to fine tune various parameters of our news recommender system, however, we argue that the success of the proposed ontology-based recommendation technique should be confirmed by a user-centred evaluation. Simulation-based evaluation fails to capture the cognitive difficulties a user faces during an information seeking process. However, as Belkin [3] pointed out in his keynote speech at ECIR 2008, bringing the user into the evaluation process is a grand challenge in the evaluation of (adaptive) information retrieval (and recommendation) approaches. In addition, there is scarsley any work evluating an information system over a longer period. Kelly [19] presents the arising challenges by highlighting that long-term evaluation of information seeking systems lack appropriate user and task models as well as test collections. Additionally, they stress the lack of multiple time series evaluation designs, a crucial problem when user satisfaction over a longer time period is used as an evaluation measure. We address these issues in this paper by introducing an interactive user-evalution methodology for the evaluation of the news recommender system.

In order to assess the recommendation approach over multiple search sessions, we introduce a between subjects multiple time series study, focusing on the following research questions:

- (RQ1) Can implicit relevance feedback techniques be exploited to create effective long-term user profiles?
- (RQ2) Can ontologies be exploited to recommend relevant news video documents?

The remainder of this paper is structured as follows. In Section 2, we introduce related work which is relevant in the context of our research. Section 3 introduces the recommender system which we employ to evaluate our research questions. Section 4 outlines the experimental design of our study. Results are discussed in Section 5. Section 6 concludes this paper.

2 Related Research

This work builds upon different research areas, including user profiling, semantic-based recommendation and interactive studies. In this section, we provide a brief introduction into these research areas.

2.1 Implicit User Profiling

User profiling is the process of learning a user's interests over a long period of time. Several approaches have been studied to capture users' news interests in a profile. Chen and Sycara [7] analyse internet users during their information seeking task and explicitly ask them to judge the relevance of the webpages they visit. Exploiting the created user profile of interest, they generate a personalised newspaper containing daily news. Bharat et al. [4] created a personalised online newspaper by unobtrusively observing the user's web-browsing behaviour. Although their system is a promising approach to release the user from providing feedback, their main research focus is on developing user interface aspects, ignoring the sophisticated retrieval issues. In the video domain, Hopfgartner et al. [17] exploit implicit relevance feedback from a community of users to recommend news video shots. Even though their approach successfully recommends video shots to given search tasks, they only consider short term content-recommendation, thus neglecting users' evolving interests that need to be captured in long-term user profiles as well as users' interests in multiple topics. In this work, we aim to address these issues by evaluating a long-term user profiling and adaptive recommendation technique.

2.2 Semantic Recommendation

A popular personalisation technique is document recommendation. Recommender systems inform users about things they might not be aware of and have not been actively searching for. They can be distinguished into two main categories: content-based recommender systems and collaborative filtering systems. Content-based recommender systems determine the relevance of an item (e.g. a video document, website or product) based on the user's interest in other, similar items. An interesting approach to identify similarity is to map relationships between concepts in the user profile by using ontologies. Fernández et al. [12] argue that ontologies can be exploited to structure news items and to annotate them with additional information. Dudev et al. [10] propose the creation of user profiles by creating knowledge graphs that model the relationship between different concepts in the Linked Open Data Cloud. This collection of ontologies unites information about many freely available different concepts. The backbone of the cloud is DBpedia, an information extraction framework which interlinks Wikipedia content with other databases on the Web such as Geonames or WordNet. In [14], we introduce an approach to identify similar news stories that match the users' interests by exploiting DBpedia. We fine tune various parameters of our model by introducing a simulation-based evaluation scheme which also highlighted the effectiveness of our approach. In this paper, we aim to confirm the outcome of this study by performing a user-centred evaluation scheme over a longer period. In the next section the challenges in the evaluation of interactive and adaptive retrieval systems are outlined.

2.3 Multiple time series studies

Christel [9] critisises that in the interactive video retrieval domain, most research approaches focus on short-term retrieval as advocated within the TRECVid evaluation campaign, hence ignoring more realistic video retrieval scenarios. Considering the broader focus of realistic video search, they argue for "Multi-dimensional In-depth Long-Term Case studies" (MILC), as advertised by Shneiderman and Plaisant [24]. Multi-dimensional stands for different evaluation measures, including interviews, surveys and logging user interaction to measure system performance. In-depth anaysis aims to include the researcher into the study process, e.g. by assisting subjects. Long-term refers to longitudinal studies, where users interact with a system over multiple sessions. Case studies aim to set the evaluation into realistic scenarios, e.g. in the users' natural environment. Long-term user studies are very common in the HCI community, but only have recently drawn the attention of the IR community. Examples include [18], which studies users' online behaviour over a time period of fourteen weeks. Every week, users were asked to fill in a questionnaire, where they had to evaluate documents they interacted with. Further, interaction logs were used to evaluate their interactions. Liu and Belkin [20] conduct a two weeks experiment where their participants were asked to perform search tasks under the authors' supervision. Thus, they evaluate long-term personalisation techniques under controlled lab conditions. Log files and questionnaires are employed as evaluation measure. Another approach is introduced by Hopfgartner and Jose [16], who outline a simulation based evaluation framework that allows for the evaluation of user modelling techniques over multiple time series. Even though their framework provides insights into the effectiveness of different recommendation techniques, they argue that a real multiple time series is required to confirm the outcome of their studies. They specify the difficulties in evaluating adaptive retrieval systems as the lack of methodologies and data collections. Lack of proper methodlogy make it difficult to control the variables involved in such a study. In addition, they argue, such approaches are very expensive to conduct in real-life. In this work, we introduce a user-centred approach to evaluating the long-term user profiles and the effectiveness of the recommendation approach.

3 System Description

In order to study the research hypotheses, a user study is required where participants use a news video recommender system over multiple days to satisfy their personal information need. It is well known that controlled experiments, i.e. experiments in a foreign environment or under someone's supervision can lead to the different behaviour of the test subjects [6]. Aiming to minimize this effect, we wanted to allow the participants of our study to perform their individual search sessions from a computer of their choice. Therefore, we implemented a Web-based news recommender system based on Asynchronous JavaScript and XML (AJAX) technology. In this section, we provide a brief overview over the system, a thorough description is given in [15].

The recommender system can be divided into different components. The first component is the video processing component that runs on a terminal equipped with a TV capture card. This component is triggered every day. The component produces an up-to-date video collection consisting of an older news corpus and latest news video stories. In this study, we focus on the daily BBC One O'Clock News and the ITV

Evening News, the UK's largest news programmes. Each bulletin has a running time of thirty minutes and is enriched with a teletext signal. Ther archive is enhanced with such daily news feeds. Following Misra et al. [21], we segment these news videos into news stories.

In the remainder of this section, we introduce the steps from annotating these news stories using external sources and indexing them. Moreover, we introduce the system interface, discuss our user profiling approach and introduce our recommendation techniques.

3.1 Semantic Annotation

Most news content providers classify their news in accordance to the IPTC standard[1], a news categorisation thesaurus developed by the International Press Telecommunications Council. We assume that a data corpus categorised by this standard will lead towards a structured user profiling approach. Therefore, following [14], we categorise news videos into the following IPTC News code subjects: Business & Finance, Disaster & Accident, Education, Entertainment & Culture, Environment, Health, Medical & Pharma, Hospitality & Recreation, Human Interest, Labour, Law & Crime, Politics, Religion & Belief, Social Issues, Sports, Technology & Internet, Weather, War & Conflict and Other. Following [14], we identify concepts that appear in the textual transcript and link these concepts with the generic ontology DBpedia. Once the link between the story and the DBpedia graph has been established, DBpedia can be exploited to put each identified entity into its context. In DBpedia, entities are nodes in a graph and a semantic hierarchy between most neighboured nodes is defined by the SKOS data model[2]. In order to identify the context of each node, we first extract all neighboured nodes in the graph which represent the category where this node belongs to. The corresponding links are defined by the property "skos:subject". Further, for each identified category node, we extract all categories that have a semantically broader meaning. These are defined by the property "skos:broader". In order to set the entities of the video stories into a broad context, we extract up to four layers of broader categories. The concept "Scotland", for example, can be set into the following hierarchy: Scotland $\subset$ United Kingdom $\subset$ European Union Member States $\subset$ European Union. Following another path, the following hierarchy can be identified: Scotland $\subset$ British Isles $\subset$ Northern Europe $\subset$ Regions of Europe.

3.2 User Interface

The interface of our system has been introduced in [13]. It can be split into three main areas: Search queries can be entered in the search panel on top, results are listed on the right side and a navigation panel is placed on the left side of the interface. When logging in, the latest news will be listed in the results panel. Search results are listed based on their relevance to the query. Since we are using a news corpus, users can re-sort their results in chronological order with latest news listed first. Each entry in the result list is visualised by an example keyframe and a text snippet of the story's transcript.

[1] `http://www.iptc.org/`

[2] `http://www.w3.org/TR/skos-reference/`

The user's interactions with the interface are expoited to identify multiple topics of interests (see Section 3.3). On the left hand side of the interface, these interests are presented by different categories. Clicking on any of these categories in the navigation panel will reveal up to four sub categories for the according category. The profiling approach will be introduced in the following section.

3.3 User Profiling

When a user interacts with a result, he leaves a "semantic fingerprint" that he is interested in the content of this item to a certain degree. In this work, we employ a *weighted story vector* approach to capture this implicit fingerprint in a profile. A detailled description of this approach is given in [15].

The content of the users' profiles is displayed on the navigation panel on the left hand side of the interface. Since the idea of such a navigation panel is to assist the users in finding other stories that match their interests, the next challenge is to identify more stories in the data corpus that might be interesting to the user.

3.4 News Video Recommendation

According to Adomavicius et al. [1], most content-based recommendation approaches identify keywords that can then be used to retrieve related content. Content-based recommendations are therefore items which have been retrieved using a personalised search query. Applying this approach, we propose to create a search query based on the content of each sub category and to retrieve stories using this query. An interesting research challenge is what should be the content of these search queries. We propose to consider the semantic link between news documents when selecting query terms.

3.4.1 Exploiting the Semantic Link

We have shown that due to the SKOS attribute, DBpedia is a directed, labelled graph D, consisting of parent and children nodes n that are labelled using globally unique identifiers (URIs). Within DBpedia, each child node can be linked to multiple parent nodes, while each parent node can be linked with multiple children nodes. Thus, $D = (V, A)$ where V is a set of nodes n and A is a set of directed edges, defined by the SKOS attribute. In general, child nodes are very specific *concepts* while parents are broader *categories*. Thus, the deeper one traverses through the knowledge graph, the more general are the parent nodes. We hypothesise that parent nodes can be exploited to identify other children nodes that are semantically related to the concepts of a news story. Imagine, for example, a user who watched a video about Scotland, but is also interested in other parts of the country. Exploiting the semantic context that Scotland is part of the United Kingdom, other videos, e.g. about England or Wales, could be recommended. Considering that both videos might not share similar terms, a purely text-based recommendation would not be able to identify the relation between both videos. We therefore propose to form personalised search queries q_n, consisting of these concepts and categories.

An interesting research question is which concepts and categories should be selected to form such queries. An obvious choice, inspired by pseudo relevance feedback

techniques, is to determine the most representative concepts. This could be, for example, the most frequent concept nodes within the sub category cluster. A problem is, however, that some documents within the cluster might contain only a small number of concept nodes. In order to overcome this sparsity, we apply a smoothing method as used in language modelling. We hence identify for every document

$$p(n|SC) = \alpha p(n|SC) + (1 - \alpha)p(n|C)$$

the probability of a concept or category n belonging to the identified subcluster SC. Thus, we also consider the appearance of each node within the whole corpus C to identify the most representative nodes. α is used as prior to smooth the impact of $p(n|C)$. In this study, we set $\alpha = 0.7$. Considering the nodes with the highest probability as representative nodes, we suggest to form a search query consisting of these nodes, combined using "or". Hence

$$q_n = n_1 \vee n_2 \vee \cdots \vee n_l, \text{ where } p(n_1|SC) < p(n_2|SC) < \ldots p(n_l|SC)$$

An important question is how many categories should be considered when formulating a search query. In [15], we identified the optimal length l of the query used to retrieve more content that is similar to the user's interest by performing a simulation based evaluation. Applying the outcome of this study, we set $l = 8$.

Another question is how the nodes, coming from different depths within the knowledge graph, should be weighted when generating the personalised search query, and how the retrieval results should be ranked. Dudev et al. [10] suggest to determine interest scores that express a user's interests in certain concepts. They argue that these scores can then be used to infer interest in related concepts. Considering the structure of the knowledge graph, we suggest to give a lower weighting to nodes from broader layers, since these nodes are rather general and consequently, their importance fades. Further, we argue that each node having the same depth should have the same weight since this allows a diverse representation of each aspect of the corresponding concept node. Thus, we define each news document in the user profile as a document composed of concept nodes and category nodes of different degrees of importance that are arranged in layers surrounding the concept nodes. We propose to formulate personalised search queries q_n consisting of l nodes n from each field, combined by "or", which have been weighted in accordance to the importance of their field. Further, similar to Pérez-Agüera et al. [22], we propose to model the semantic structure by ranking the results using BM25F [26]. Thus, we model the decreasing importance of more broader concepts by giving a lineary higher weight to lower concepts.

3.4.2 Text-based Recommendation

In [15], we introduced two text-based search queries, consisting of Named Entities and Nouns/Foreign Names, that can be employed as a potential Baseline Recommendation Component. The results of the simulation runs suggest that Named Entities result in better recommendations than Nouns and Foreign Names. Consequently, we rely on Named Entities in our Baseline System. Similar to our DBpedia recommendation, we determine

$$p(e|SC) = \alpha p(e|SC) + (1 - \alpha)p(e|C)$$

the probability of a named entity e belonging to the identified sub category cluster SC and suggest to form a search query consisting of these entities, combined using "or". Thus,

$$q_e = e_1 \vee e_2 \vee \cdots \vee e_l, \text{ where } p(e_1|SC) < p(e_2|SC) < \ldots p(e_l|SC)$$

A search query q_e hence consists of l highest ranked named entities. Aiming to ease comparison of our approaches, we rank results using the classical ranking function Okapi BM25.

4 Experimental Methodology

Aiming to assess the recommendation approach over multiple search sessions, we performed a between subjects multiple time series study. Participants, recruited using various mailing lists, were paid a sum of £15 to take part in our evaluation.

The experiment started with a short introduction in the experimenter's office, where the participants got familiarised with the experiment. After filling an Entry Questionnaire we introduced them to the news video recommender system in a 10 minute training session.

We split the experiment into ten sessions that the participants could perform from a computer of their choice. They were asked to include the news recommender system in their daily news consumption routine and interact with the system for a minimum of 10 minutes each day. Half of them, randomly assigned, were asked to interact with the Baseline system while the other half used the Semantic system. Every day, they received an email that reminded them to continue with the experiment. The participants were told to use the system to explore any news they are interested in. Further, as proposed by Borlund [5], we created a simulated search task situation that they could search for in case they did not find any other news of interest. Our expectation was twofold: First of all, we wanted to guarantee that every user had at least one topic to search for. Moreover, we wanted the participants to actually explore the data corpus. Various sport events happened during the time of the experiment. We therefore chose a sports dominated scenario:

> *"You just started a new job in a remote town away from home. Since you do not know anyone in town yet, you are keen to socialise with your colleagues. They seem to be big sports supporters and are always up for some sports related small talk. Sport hence opens great opportunities to start conversations with them. Luckily, there are various major sports events and tournaments this month which they all show interest in, e.g. the Winter Olym-pics in Vancouver, the Rugby Six Nations Cup and European Football Tournaments. Every day, you eagerly follow every news report to be up to date and to join their conversations."*
>
> **Indicative Request: You should use the recommender system to follow sports related news. This might include major international events such as the Winter Olympics, European football competitions or the Rugby Six Nations cup. Reports might be summaries of the competition day, feature reports about Team GB, or summaries of football/rugby matches. Keep in mind that you should follow the news well enough to be able to chat and socialise with your new colleagues.**

We wanted to evaluate the recommendation approaches by comparing the participant's opinions about the systems during various stages of the experiment. Campbell and Stanley [6] advise not to repetitively prompt the same questions within a short time period, since users' behaviour may adapt based on the intention of the questionnaires. Therefore, we asked used to fill a questionnaire them every second day of the experiment. At the end of their tenth search session, they were asked to fill in an online Exit Questionnaire.

Note that due to the uncontrolled nature of the experimental setting, we had no influence on when and how the participants performed their search sessions. Consequently, the experiment lasted roughly one month, since some participants skipped various dates, hence increasing the overall duration of the experiment. Nevertheless, even though this was not intended, we argue that such user behaviour resulted in more realistic data. Most users check news whenever they feel, and not necessarily every day, as initially intended by us. Further, we argue that skipping days within the experiment can be an interesting challenge for our user profiling approach.

4.1 Data Corpus

The news video recommendation approach relies on an up-to-date news video corpus which will be updated constantly. Prior to starting the experiment, we recorded the daily news broadcasts from BBC One (One O'Clock News) and ITV (Evening News) for various months and processed the bulletins as outlined before. During the experiment, we automatically updated the corpus by recording and processing the latest news broadcasts from these channels. The participants could hence explore the latest news and access older news. The following events were scheduled at the time of the experiment: The XXI Olympic Winter Games, the first leg of the Round of 16 of the UEFA Champions League, the first and second lef of the Round of 32 of the UEFA Europa League and the RBS Six Nations Rugby Championship.

The reported events provide different conditions for our user profiling and recommendation techniques. The Winter Olympics, for example, took place on every day during the experiment. The profile of a user showing interest in the Games would consequently contain many news stories about the Games. The football and rugby games were less frequent. A user interested in these games would hence have less interaction with corresponding reports.

4.2 Participants

24 users (16 male and 8 female) participated in the experiment. 21 of them were either Graduate Students or Faculty/Research Staff, three hold an undergraduate degree. The majority of the participants studied or worked on Computer Science related topics, mostly focusing on Human Computer Interaction and Information Retrieval. They claimed to have a high expertise of English with 25% of them being native speakers of English. The majority (62%) of the participants were between 26–30 years old. 24% were between 31–38 years old and three participants were in the age group of 18–25 years.

5 Results

In order to evaluate the previously introduced research questions, we followed a user-centred evaluation scheme where the users' satisfaction and interaction are the most valuable evaluation measures. By asking for frequent reports every second day of the study, our goal was to evaluate the users' opinion about the system at various stages of the experiment. As suggested by Ruthven and Kelly [23], this enables us to get a better understanding of the performance of the recommendation approaches over a longer time period. Further, tracking user interactions in log files allows us to get an insight into user activities. A mixed between-within subjects analysis of variance was conducted to assess the participants' answers across the different stages of the experiument. The main effect comparing the two systems was significant, Wilks' Lambda = 0.808, F = 3.069, p = 0.004, suggesting a difference in the effectiveness of the two approaches. Additional statistical significance tests are performed on every differential of these questionnaires. In this section, we present a more detailed analysis of these questionnaires and the created log files, including a report on all statistically significant results.

5.1 System Usage and Usability

The first question of our interim questionnaire was to find out what the participants actually used the system for. We therefore asked them to check on the online form those pre-defined tasks that described best their activity. The majority of participants used it to retrieve the latest news, followed by identifying news stories they were not aware of before. Furthermore, we were curious to see what news categories they were interested in. The participants were therefore asked to check on the questionnaire the corresponding news categories they were interested in during the last days.

Table 1 News categories that the users were interested in during the experiment

	Total	Percentage
Business & Finance	29	25%
Entertainment & Culture	42	36%
Health, Medical & Pharma	26	22%
Politics	62	53%
Sports	78	67%
Technology & Internet	33	28%
Other	8	7%

As Table 1 indicates, the participants followed various news categories. Note that people could select more than one checkbox. Thus, percentages add up to more than 100%. These diverse answers suggest that users did not only use the system to retrieve stories of the pre-defined search task, but also used it for their own information needs, i.e. to follow latest news or to discover other news stories that match their interests.

In order to evaluate the users' satisfaction while interacting with the interface, we asked the participants to judge various statements on a Five-Point-Likert scale from 1 (Agree) to 5 (Disagree). The order of the agreements varied over the questionnaire to reduce bias. In order to determine the general usability of the system, we asked them to judge the following statements: (1) "The interface structure helped me to explore

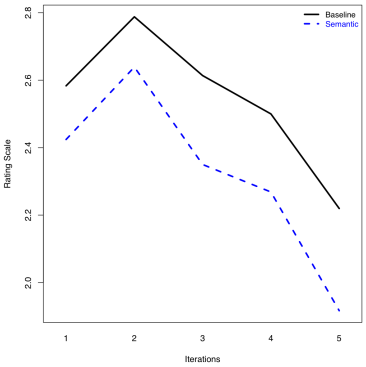

Fig. 1 The interface helped me to explore the news collection (lower is better)

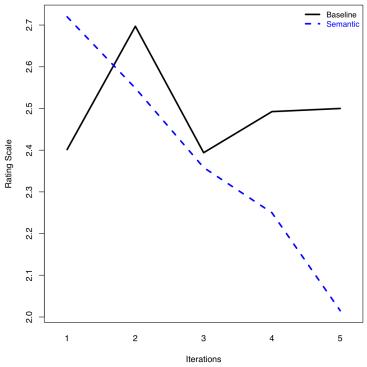

Fig. 2 The interface helped me to explore various topics of interest (lower is better)

the news collection" and (2) "The interface helped me to explore various topics of interest". Figures 1 and 2 show the average judgements of all users over all ten days.

Interestingly, the two different user groups had a different perception of the accessability of the collection and topics of interest. Considering that both groups interacted with the same interface, we assume that the participants generalised their judgments with respect to the whole system they used rather than the interface only. Figure 1 shows the users' agreement that the system helped them to explore the news collection. Neglecting a bump at the second iteration, i.e., the fourth day of the study, a clear trend towards positive perception can be observed. This trend can be explained by humans' learning abilities. Once the users got used to the system, they appreciated its functionalities. The bump at the second iteration might be explained by a bug that occurred during the fourth day of the study: even though categories and search results were displayed, the interface did not allow the users to access any sub categories. The same bug can explain the bump that is shown in Figure 2, depicting the users' opinion about the systems' usability to explore various topics of interest. Remarkable is the better assessment of the semantic based system which suggests an overall better performance of this recommendation technique. A two way analysis of variance (ANOVA) across both systems and the different iterations revealed that the users' perception of the novelty of provided documents were dependent ($p = 0.0025$).

5.2 Exploiting Implicit Relevance Feedback

Figure 3 provides an overview over the average number of implicit relevance feedback that the participants provided while interacting with the system. The figures, extracted from the log files of the study, illustrate that users performed a vast amount of interactions that could be used to identify their interests. Note that overall, the user group interacting with the semantic recommender system provided more implicit relevance feedback than the users of the baseline system. Moreover, a high activity can be spotted on the first day of the study by the semantic user group. A closer look at the log files reveals that one user browsed through all key frames of the news stories he retrieved. Considering that he performed this action on the first day only, we consider this as an

anomaly. Another interesting observation is the decreasing amount of feedback that the users of the baseline system provided at later stages of the experiment. This could indicate that they lost interest in the study, maybe due to the inefficiency of provided recommendations.

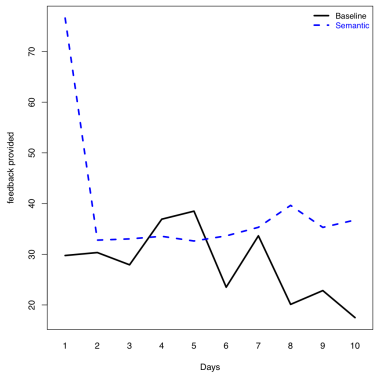

Fig. 3 Implicit relevance provided by both user groups

With the aim of studying our first research question (RQ1) whether implicit relevance feedback can be used to create long-term user profiles, we were interested whether the system was effective in automatically identifying the users' interests. Therefore, we asked the participants to judge the following statement: "Categories were successfully identified and displayed on the left hand side of the interface". Their judgements are depicted in Figure 4. Further, we aimed to understand whether the content of these categories matched the users' interests. Figure 5 illustrate their judgements of the statement "The displayed sub categories represent my diverse interests in various topics".

As can be seen, the users of the semantic based recommender system provided a better assessment than the users of the baseline system. It can be seen that the ini-

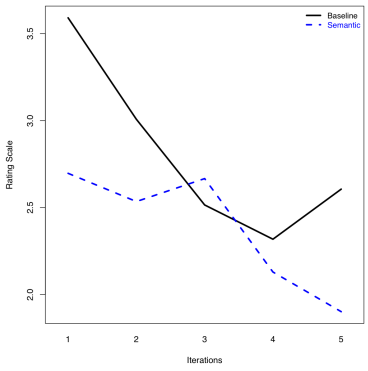

Fig. 4 Categories were successfully identified and displayed on the left hand side of the interface (lower is better)

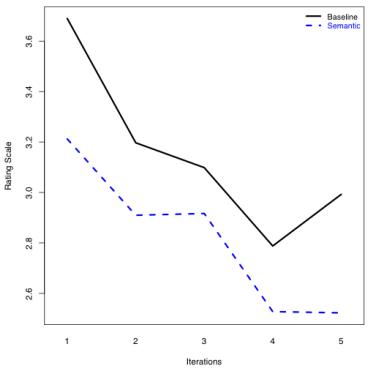

Fig. 5 The displayed sub categories represent my diverse interests in various topics. (lower is better)

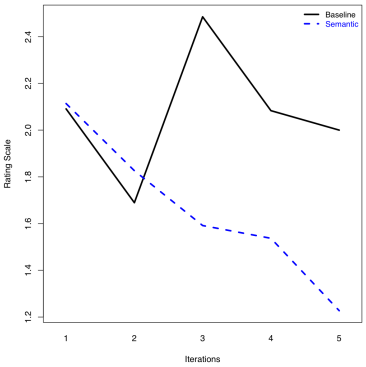

Fig. 6 The displayed sub categories represent my diverse interests in various topics. (lower is better)

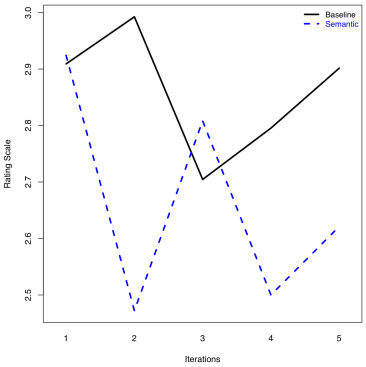

Fig. 7 The displayed results for each sub category were related to each other. (lower is better)

tial assessment in the first iteration is rather negative, i.e., above the mean of three. Considering that profiling approaches relies on preceding user input, this weak assessment can be explained by the "cold start phenomena": Without any user feedback, the system cannot identify users interests. Both groups, however, developed a more positive perception, hence indicating that at later stages of the experiment, the displayed sub categories became more focused. The participants statement that relevant categories where identified were dependent on the system ($p < 0.0047$ for significance). Summarising, both questions suggest that implicit relevance feedback can be used to create long-term user profiling, thus answering our first research question.

Further, we asked users to judge the following two statements: (1) "The displayed sub categories represent my diverse interests in various topics" and (2) "the displayed results for each sub category were related to each other". Figures 6 and 7 show the average answers over the whole time of the experiment. As can be seen, the average ratings from both groups indicate a positive tendency towards the two statements. Thus, their responses suggest that implicit relevance feedback can be used to capture users' long-term interests. Note, however, that the agreement is rather fluctuant. Moreover, a two way ANOVA did not reveal a significant difference between both groups for the above introduced differentials.

5.3 Recommendation Quality

Finally, aiming to evaluate the second research question (RQ2) whether ontologies can be used to recommend news videos, we asked the participants to judge whether "the displayed results for each category contained relevant stories [they] did not retrieve otherwise". This statement aimed to evaluate whether the recommendations provided some novelty to the user and did not just consist of news documents that the users had seen before already. Further, we asked them to assess the statement "the displayed results for each category matched with the category description". With this statement, we aimed to understand whether the recommendations were in the right context. The participants' replies are depicted in Figures 8 and 9.

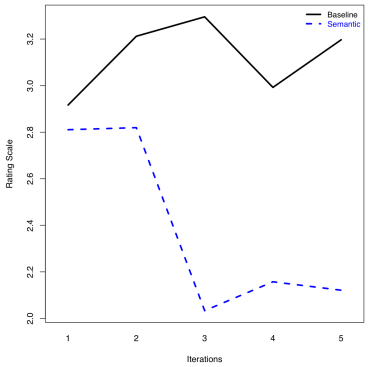

Fig. 8 The displayed results for each category contained relevant stories I didn't receive otherwise. (lower is better)

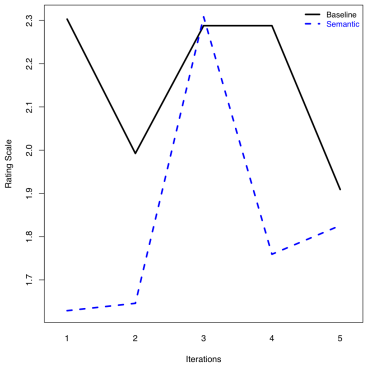

Fig. 9 The displayed results for each category matched with the category description. (lower is better)

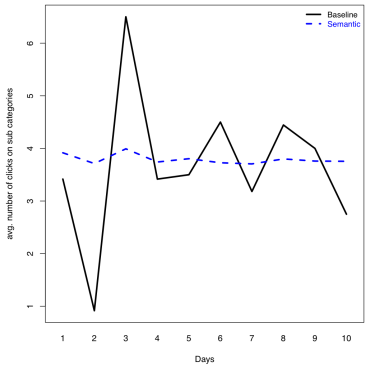

Fig. 10 The average number of clicks on the sub categories

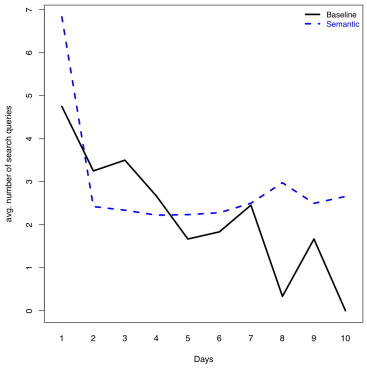

Fig. 11 The average number of manually triggered searches

Note that even though no significance can be reported, the semantic-based recommendation run received an overall higher weighting than the baseline run. Thus, the results support the outcome of our simulation that ontologies can be successfully employed to provide news video recommendations. Aiming to evaluate this observation, we further analysed the users' search behaviour. Figures 10 and 11 show the average number of recommendations that were triggered by the users, i.e. the average number of clicks on the sub categories and the average number of manually triggered searches, respectively.

While users of the semantic recommender system constantly triggered recommendations over all days of the experiment, Figure 10 shows a less homogeneous interaction pattern of the baseline system's user group. Moreover, Figure 11 shows a decline in the amount of manually triggered search queries of the baseline group, while the second group constantly triggered their own searches. Both observations could indicate the dissatisfaction that the users of the baseline system experienced during the study. The longer they interacted with the system, the less convinced they seemed to be to use it to retrieve videos. Moreover, the irregular usage of the recommendations might

indicate that the users interacted with the system rather randomly, maybe just to fulfil the search task. Given that no statistical significance can be reported, a more thorough analysis is required to confirm the outcome of the simulation-based evaluation reported in [15] that ontologies can be exploited to recommend relevant news documents

6 Conclusion

In this paper, we aimed to confirm the outcome of the simulation-based evaluation of long-term news video recommendation that has been outlined in [15] by employing a multi-session time-series user study. We focused on two research questions: In the first research question (RQ1), we aimed to study whether implicit relevance feedback techniques can be exploited to create efficient long-term user profiles. Implicit relevance feedback is a well established tool to identify short-term users' interests with interrupting their search process (e.g., [4,7,17,25]). The results of our simulation-based study [15] suggest that the same technique can be employed to create long-term user profiles. In the second research question (RQ2), we studied whether ontologies can be exploited to recommend relevant news video documents. As we have shown, many studies have been performed to exploit semantic technologies for recommending relevant documents (e.g., [10,12]). Differing from our approach, these studies mainly focus on short-term recommendations, thus neglecting long-term recommendation techniques.

In order to address these issues, we suggested an experimental methodology where 24 users were asked to include an online news video recommender system over a time period of ten days into their daily news gathering routine. The news recommender system automatically captures daily broadcasting news and segments the bulletins into coherent news stories. We evaluated two types of recommendations that have been evaluated and fine-tuned in the previous chapter. Differing from standard interactive information retrieval experiments, this evaluation was split into multiple sessions and was performed under an uncontrolled environment, two necessary conditions for a realistic evaluation of implicit user profiling. This novel approach cannot rely on system-centred evaluation measures as common in information retrieval experiments. Thus, standardised evaluation measures need yet to be developed.

We evaluated the two introduced research questions by analysing both the questionnaires and log files of our user study. As we have shown before, the participants were asked to express their opinion about the recommender system by filling in interim questionnaires at various stages of the user study.We analysed these questionnaires under three different criteria. In general, we aimed to evaluate the general usability of the system and the way that users interacted with it. The participants' responses indicate that they used the recommender system to explore various types of news topics and that they found the system very helpful. Our analysis results show a statistically significant preference for the semantic-based system, thus answering the research questions (Section 5). This observation can be backed up by a statement that one participant formulated in the exit questionnaire of the experiment: "In general, the system is great to explore news according to the user interests. The automatic organization of topics and interest evolved through the time I used the system, and I did not need to search again using the keyword box. That was definitely nice." Summarising, we conclude that the introduced system has the potential to improve users' news gathering routines. As shown in Figures 811, this verbal feedback is supported by the data mined from the log files as well as questionnaire data. Further, we discussed various questions that aimed to

evaluate whether implicit relevance feedback can be used to capture users' interest over a longer period (RQ1). We compared the amount of categories that were successfully identified and displayed on the interface when using the baseline and semantic-based technique, correspondingly. Further, we analyzed the users' opinion whether the system was able to capture their interests in different aspects of news. The users' responses suggest that the introduced technique successfully captured users' broad interests and was able to successfully identify sub interests. Statistical significance was reported. Thus, we argue that implicit relevance feedback can be employed to create long-term user profiles.

Finally, we introduced the users' judgements about the recommendation quality, aiming to study (RQ2) whether ontologies can be exploited to recommend relevant news video documents. Positive tendencies could be spotted by both user groups, with the semantic-based recommender system achieving a better rating than the baseline system. Statistically signifiant difference cannot be reported though. Nevertheless, we conclude that our semantic-based recommendation technique can successfully be employed to provide novel and relevant news recommendations over a longer time period. A more thorough analysis of the role of ontologies to study is required to draw final conclusions regarding (RQ2). Further work includes studying, which implicit indicators shall be considered as positive and negative feedback, respectively. Further, a weighting should be determined that expresses the users' interests appropriately.

References

1. G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. on Knowl. and Data Eng.*, 17(6):734–749, 2005.
2. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. DBpedia: A Nucleus for a Web of Open Data. In *Proc. 6th Int. Semantic Web Conf.*, pages 722–735. Springer Berlin / Heidelberg, 11 2007.
3. N. J. Belkin. Some(what) grand challenges for information retrieval. *SIGIR Forum*, 42(1):47–54, 2008.
4. K. Bharat, T. Kamba, and M. Albers. Personalized, interactive news on the web. *Mult. Syst.*, 6(5):349–358, 1998.
5. P. Borlund. The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3), 2003.
6. D. T. Campbell and J. C. Stanley. *Experimental and Quasi-Experimental Design for Research*. Wadsworth Publishing, Monterey, CA, 1 edition, 1963.
7. L. Chen and K. Sycara. WebMate: A personal agent for browsing and searching. In K. P. Sycara and M. Wooldridge, editors, *Proc. Agents'98*, pages 132–139, New York, 9–13, 1998. ACM Press.
8. M. G. Christel. Establishing the utility of non-text search for news video retrieval with real world users. In *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia*, pages 707–716, New York, NY, USA, 2007. ACM.
9. M. G. Christel. Examining user interactions with video retrieval systems. In *SPIE'06: Proceedings of SPIE Volume 6506, Multimedia Content Access: Algorithms and Systems*, 2007.
10. M. Dudev, S. Elbassuoni, J. Luxenburger, M. Ramanath, and G. Weikum. Personalizing the Search for Knowledge. In *Proc. PersDB*, 08 2008.
11. M. Fernández, V. López, M. Sabou, V. Uren, D. Vallet, E. Motta, and P. Castells. Using TREC for cross-comparison between classic IR and ontology-based search models at a Web scale. In *SemSearch'09*, 4 2009.
12. N. Fernández, J. M. Blázquez, J. A. Fisteus, L. Sánchez, M. Sintek, A. Bernardi, M. Fuentes, A. Marrara, and Z. Ben-Asher. NEWS: Bringing Semantic Web Technologies into News Agencies. In *Proc. ISWC*, pages 778–791, 2006.

13. F. Hopfgartner. Adaptive interactive news video recommendation: An Example System. In *Proc. of SEMAIS'11 - Second International Workshop on Semantic Models for Adaptive Interactive Systems*, pages 21–25, 01 2011.
14. F. Hopfgartner and J. M. Jose. Semantic User Modelling for Personal News Video Retrieval. In *Proc. of the Multimedia Modelling Conference*, pages 336–349, 01 2010.
15. F. Hopfgartner and J. M. Jose. Semantic user profiling techniques for personalised multimedia recommendation. *ACM/Springer Mult. Syst.*, 16(4):225–274, 2010.
16. F. Hopfgartner and J. M. Jose. Development of a test collection for studying long-term user modelling. In *Proc. of LWA*, 2011.
17. F. Hopfgartner, D. Vallet, M. Halvey, and J. M. Jose. Search Trails using User Feedback to Improve Video Search. In *Proc. of the ACM Int. Conf. on Multimedia*, pages 339–348, 10 2008.
18. D. Kelly. *Understanding implicit feedback and document preference: A naturalistic user study*. PhD thesis, Rutgers University, 2004.
19. D. Kelly, S. T. Dumais, and J. O. Pedersen. Evaluation Challenges and Directions for Information-Seeking Support Systems. *IEEE Computer*, 42(3):60–66, 2009.
20. J. Liu and N. J. Belkin. Personalizing information retrieval for multi-session tasks: The roles of task stage and task type. In *SIGIR'10*, pages 26–33, 2010.
21. H. Misra, F. Hopfgartner, A. Goyal, P. Punitha, and J. M. Jose. News video story segmentation based on semantic coherence and content similarity. In *Proc. of Multimedia Modeling Conference*, pages 347–357, 01 2010.
22. J. R. Pérez-Agüera, J. Arroyo, J. Greenberg, J. Perez Iglesias, and V. Fresno. Using bm25f for semantic search. In *Semantic Search 2010 Workshop*, 2010.
23. I. Ruthven and D. Kelly. Interactive information seeking behaviour and retrieval. Facet Press, 2010.
24. B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In *BELIV'06*, pages 1–7, 2006.
25. D. Vallet, F. Hopfgartner, J. M. Jose, and P. Castells. Effects of usage-based feedback on video retrieval: a simulation-based study. *ACM Trains Inf Syst*, 29(2):11:1–11:32, 2011.
26. H. Zaragoza, N. Craswell, M. J. Taylor, S. Saria, and S. E. Robertson. Microsoft Cambridge at TREC 13: Web and Hard Tracks. In *TREC'04: Proceedings of the Thirteenth Text REtrieval Conference*, 11 2004.